강우원, 김다윤, 김소연, 정용헌

NLP 감정분석 기반
마케팅 시장 분석

NLP 감성 분석에 기반한 온라인 커뮤니티 이슈 키워드 모니터링 시스템

빅데이터미네이터

“ ☹😐☺

커뮤니티별 주요 이슈 키워드,
감성 분석 시스템

각 커뮤니티의 제목, 본문, 댓글에 대한 크롤링을 진행하고, 이에 따른 주요 이슈 키워드를 추출한다.

또한 추출한 데이터를 바탕으로 감성 분석하여 결과를 도출하고자 한다.

커뮤니티별 익일 이슈 키워드와 그에 대한 감성 분석 결과를 사용자에게 알려주는 시스템이다.

최종 목표는 매일 온라인 커뮤니티의 주요 이슈 키워드와 그에 대한 감성 분석을

사용자에게 알람으로 전달해주는 서비스이다.

따라서 선정한 커뮤니티인 4곳의

인기 게시판의 하루 치 데이터를 크롤링해 test 데이터를 만들었다.

이를 바탕으로 각 커뮤니티에대한 키워드 추출, 감성 분석 등을 진행해 나갔다.

키워드 추출

▪ 명사 추출, Bag of Words, TF- IDF, 토픽 모델링 등 여러 방식을 조사해보았다. 또한 의미 있는 키워드 추출을 위해 한가지 방식만 적용하는 것이 아닌, 2가지 방식을 결합해서도 시도해 보았다.

결과적으로는 명사 추출을 한 다음 TF-IDF 하는 방식이 프로토타입 단계에서 가장 활용 가능한 결과를 내는 것으로 판단하였다.

In [5]:
```
#TF-IDF 이미 한번 명사 단위로 정리 한 명단으로 처리할 시

as_one = ''
for noun in nouns:
    as_one = as_one + ' ' + noun
words = as_one.split()

counts = Counter(words)

vocab = sorted(counts, key=counts.get, reverse=True)

#단어 빈도 정리
word2idx = {word.encode("utf8").decode("utf8"): ii for ii, word in enumerate(vocab,1)}

#딕셔너리로 정리
idx2word = {ii: word for ii, word in enumerate(vocab)}
```

In [6]:
```
#tf-idf
tfidf = TfidfVectorizer(max_features = 10, max_df=0.95, min_df=0)

#generate tf-idf term-document matrix
A_tfidf_sp = tfidf.fit_transform(nouns)  #size D x V
```

In [7]:
```
#tf-idf dictionary
tfidf_dict = tfidf.get_feature_names()
print(tfidf_dict)

morphs_dict = {
    '키워드' : tfidf_dict,
}

df2 = pd.DataFrame(morphs_dict)
df2.to_csv('10키워드s.csv', index=False, encoding="utf-8-sig")
```

```
['결혼', '남자', '남편', '본인', '부모', '시간', '엄마', '여자', '자기', '친구']
```

감성 분석: 01. 네이트판

1. 키워드 Top 10 추출

```
['결혼', '남자', '남편', '본인', '부모', '시간', '엄마', '여자', '자기', '친구']
```

2. 1위 키워드에 대한 데이터만 추출 및 토큰화

	Keyword	content	positive	negative	neutral
0	결혼	['엄', '시', '어머니', '순간', '시', '어머니', '정', '엄마',...	0	0	0

KNU 감성어 사전 활용하여 감성 분석

	Keyword	negative	neutral	positive
0	결혼	-397	13	293

⇒ 앞으로 수정해 나가야하는 것

1) '예랑' 등 인터넷 용어와 '온도차', '인성' 등 사전에 없지만 온라인에서 많이 사용되는 감성어를 앞으로 계속해서 사전에 추가해 나가야 한다.

2) 데이터 전처리시, 정확하지 못한 띄어쓰기로 인한 오류를 해결하기 위해, 현재 시도해본, 공백 모두 지우고 pykospacing으로 띄어쓰기 다시 해주는 방식 외에 더 효과적인 방식이 있는지 조사 해볼 것이다.

감성 분석: 02. FM

1. 키워드 Top 10 추출

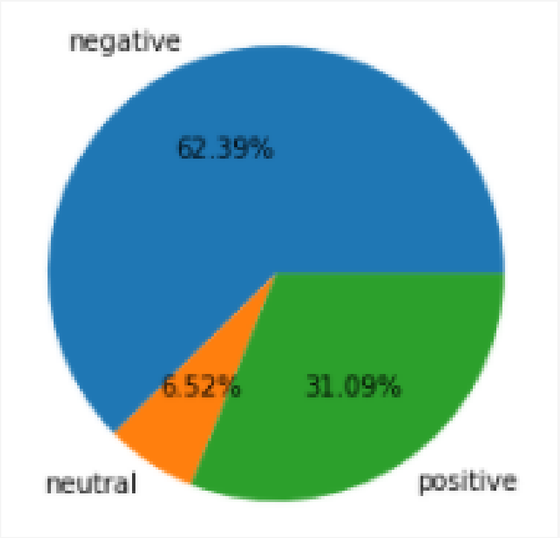
```
['간부', '군대', '기부', '나라', '백신', '병원', '수술', '아웃', '추신수', '후원']
```

2. 1위 키워드에 대한 데이터만 추출 및 토큰화

```
'거부', '어머니', '전화', '코로나', '규정', '인하다', '민간', '병원', '수술', '절대', '받다', '전역', '수술',
'받다', '하다', '전함', '쓸다', '고통', '견디다', '복무', '계속', '하다', '마비', '통증', '점점', '심해지다',
'차례', '수술', '필요하다', '소견', '제출', '하다', '휴가', '요청', '하다', '민간', '병원', '예약', '되다',
'휴가', '불가능하다', '까지', '거절', '쓸다', '부모님', '동해', '예약', '하다', '하다', '사진', '만으로는',
'예약', '불가능하다', '결국', '휴가', '감수', '병원', '에서', '진료', '받다', '지다', '개월', '지나', '전역',
'하다', '추다', '휴가', '나오다', '응급', '수술', '받다', '후유증', '지체', '장애인', '되다', '버리다', '쓸
다', '개월', '만원', '재활', '치료', '간병인', '까지', '써다', '되다', '상황', '이지만', '전역', '하다', '이
유', '군대', '에서', '하다', '보상', '받다', '있다', '십지어', '간부', '에게', '사과', '조차', '받다', '끝나
다', '버리다', '생활', '찾다', '국가', '아들', '다치다', '아들', '군대', '갈다', 'ㅁ', '좆병', '신다', '들
다', '살인', '마렵다', '왤다', '군대', '쓰레기', '꼬이다', '임자', '복무', '진급', '문제', '생기다', '되다',
'되다', '전역', '하다', '전역', '하다', '책임', '간부', '쏘다', '죽이다', '원래', '민간', '병원', '수술', '치
료', '받다', '자다', '보다', '코로나', '핑계', '민간', '수술', '막다', '문제', '간부', '전쟁', '나다', '대가
```

Polarity 감성 사전 활용하여 감성 분석

	부정	중립	긍정
0	244.099291	25.490231	121.650713



⇒ 앞으로 수정해 나가야하는 것

1) KNU 감성 사전과 Polarity 감성 사전 중 어느 것이 프로젝트에 더 적합한지 비교할 것이다.

2) 토큰화할 때 동사까지 의미 없이 토큰화 되어 버리는 경우가 있어서, 더 나은 토큰화 방법과 데이터 정제 방법을 찾아볼 것이다. 예) 하고싶다 → 하고/싶다

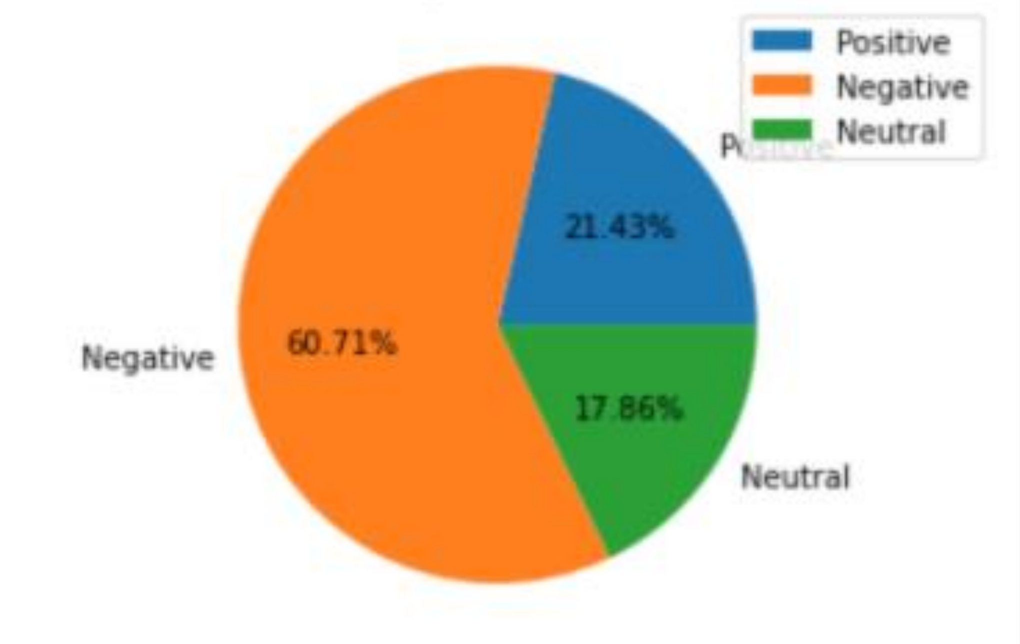
감성 분석: 03. 루리웹

1. 키워드 Top 10 추출

['매수', '민주당', '부모', '사건', '새끼', '유튜브', '의대', '자식', '친구', '한강']

2. “한강“ 키워드 : KNU 감성어 사전 활용하여 감성 분석

	sentence	comment	score	ratio
0	난 한강사건 초기부터 지금까지 느끼는건 딱 하나야,대체왜,우리나라 터졌을때 문재인...	[난, 한강, 사건, 초기, 부터, 지금, 까지, 느끼는건, 딱, 하나, 야, 대체...	-2.0	-0.63
1	한강 사건을 보면서 느낀 두 가지,힘드네 어쩌네 그런거 다 엄살처먹고 살만하니까 방...	[한강, 사건, 을, 보면서, 느낀, 두, 가지, 힘드네, 어쩌, 네, 그런거, 다...	-1.0	-0.32
2	어제자 그알 보고 느낀 점,코로나가 염병이긴 염병이구나코로나로 온 동네 노인정 다...	[어제, 자, 그, 알, 보고, 느낀, 점, 코로나, 가, 염병, 이긴, 염병, 이...	2.0	0.64
3	그알이 남긴 역사로 남을 짤,,조선일보도 못믿는다더라 탄핵시국에 통수쳤다고 촌의 뒤...	[그, 알, 이, 남긴, 역사, 로, 남, 을, 짤, 조선일보, 도, 못, 믿는다더...	4.0	1.25
4	한강 코난들이 ↗됐다고 들었습니다,이제 찻값 받으셔야지 ,내가 봐서는 ↗되지 않았음...	[한강, 코난, 들, 이, 됐다고, 들었습니다, 이제, 찻값, 받으셔야지, 네, 가...	0.0	0.00
5	한강사건은 진짜 검경수사권 조정 되었으니 다행이지,안그랬으면 씨 집안이랑 청와대 ...	[한강, 사건, 은, 진짜, 검, 경, 수사권, 조정, 되었으니, 다행, 이, 자안...	-2.0	-0.63
6	한강사건베글에 한강사건 그알 인터뷰 나온거 보는데,엔딩에 씨 가족내가 보기에는 씨 ...	[한강, 사건, 베글, 에, 한강, 사건, 그, 알, 인터뷰, 나온거, 보는데, 엔...	-4.0	-1.19
7	그알이 비겁한게 언레기는 안 다루더라,언레기 책임은 안 다루고 전부 유튜버로 몰아가...	[그, 알, 이, 비겁한게, 언레기, 는, 안, 다루더라, 언, 레기, 책임, 은,...	-1.0	-0.27
8	한강 코난들 됨음 ,로펌에서 고소하기로 함 ,인터넷 글 흔적없이 박제걱정 없이 삭제...	[한강, 코난, 들, 됨음, 로펌, 에서, 고소하기로, 함, 인터넷, 글, 흔적, ...	-3.0	-0.93
9	한강사건 친구 허위사실유포 자료 수집중. 요즘은 인성도 능력이지 능력주의 신봉자들 ...	[한강, 사건, 친구, 허위, 사실, 유포, 자료, 수, 집중, 요즘, 은, 인성, ...	1.0	0.32



감성 분석: 04. 클리앙

1. 키워드 Top 10 추출

```
==============================
[('독보적인', 0.5773502691896257), ('포지션의', 0.5773502691896257), ('라면jpg', 0.5773502691896257)]
[('곡', 0.0), ('해야', 0.7071067811865475), ('합니다', 0.7071067811865475)]
[('대통령', 0.5505723023275709), ('사저', 0.8347874818836936)]
[]
[]
[('시계', 0.447213595499958), ('보도했던', 0.447213595499958), ('SBS', 0.0), ('이승재', 0.447213595499958), ('기자', 0.4
47213595499958)]
[]
[('성과가', 0.7071067811865475), ('뭔가요', 0.7071067811865475)]
[('전창관', 0.5773502691896257), ('메북통하가는', 0.5773502691896257), ('길', 0.0), ('입니다', 0.5773502691896257)]
[('사진관', 0.3356360527834813), ('문재인', 0.3033829386533581), ('대통령', 0.2792927699687521), ('그레고리', 0.4234686
473781607), ('추기경을', 0.4234686473781607), ('뵙고', 0.4234686473781607), ('왔습니다', 0.4234686473781607)]
[('터트리게', 0.5), ('하는', 0.5), ('중앙', 0.5), ('헛소리', 0.5)]
[('호건', 0.408248290463863), ('메릴랜드', 0.408248290463863), ('주지사', 0.408248290463863), ('"추모의', 0.0), ('벽"
0.0), ('25만불', 0.408248290463863), ('지원', 0.408248290463863)]
```

2. 감성 분석

```
==============
1
1
1
0
1
0
1
2
1
1
1
0
2
1
0
2
2
```

```
정부jpg : 부정
아닌가봅니다 : 부정
업데이트 : 긍정
남페미였네요 : 부정
주장 : 부정
그림 : 긍정
만나다 : 부정
근황jpg : 부정
ㅋㅋ : 부정
이유jpg : 긍정
안드나요 : 부정
해체되네요 : 부정
방 : 긍정
염려 : 부정
SNSjpg : 부정
열린공감TV : 부정
```

+

최종 발표 전까지

1. 전처리 수정

: 다양한 전처리 방식 계속 시도해보고, 더 성능이 나은 방식 채택

2. 감성어 사전에 단어 추가

: 감성어 사전에 현재 없는 온라인 용어 사전에 계속 추가

3. 전체적인 코드 정리 및 수정

- 감성 분석 방식 채택

- 키워드 10개에 전체에 대한 감성 분석

감사합니다 :)

빅데이터미네이터

손윤식 교수님

주식회사 닥터송 송환구 멘토님